Proc. Nat. Acad. Sci. USA
Vol. 69, No. 7, pp. 1820–1824, July 1972

Patterns of Genetic Differentiation in the Slender Wild Oat Species *Avena barbata*

(California/Mediterranean/natural selection/allozyme polymorphisms/ morphological polymorphisms/electrophoresis)

M. T. CLEGG AND R. W. ALLARD

Department of Genetics, University of California, Davis, Calif. 95616

Communicated by Theodosius Dobzhansky, May 1, 1972

ABSTRACT **Allozyme frequencies at five enzyme loci were determined for 14 California populations of *Avena barbata*, a species introduced to California from the Mediterranean Basin during the colonization of North America. Allelic frequencies at these loci were also determined in Mediterranean collections of this species. The pattern of divergence of the California populations from the ancestral gene pool was not random and was strongly correlated with environment; thus, the pattern is not in accord with the hypothesis that most electrophoretically detectable variants are adaptively neutral. Rates of gene substitution in California were also not in accord with the neutrality hypothesis. The observations are, however, compatible with predictions of Neo-Darwinian evolutionary theory. We interpret these observations to indicate that natural selection plays a major role in determining the unique patterns of distribution of genetic variability in the slender wild oat in California.**

A substantial body of evidence indicates that organisms as different as man and predominantly self-pollinating species of plants are extensively polymorphic for a high proportion of genetic loci that govern protein variants detectable by electrophoresis (1–5). The existence of such large amounts of genetic variability is inconsistent with some population models because, according to these models, the maintenance of so much genetic variability by selection would impose intolerable genetic loads on populations. To circumvent this problem, proponents of such models have proposed that a considerable proportion of amino-acid substitutions in proteins are irrelevant to function and, hence, do not contribute to genetic load because they are selectively neutral (6–8).

The neutrality hypothesis leads, fortunately, to several quantifiable predictions that relate mutation rates and migration rates to effective population size and, thus, to expected patterns of allelic distribution among local populations, and also to evolutionary rates (9, 10). In particular, neutrality models lead to the prediction that frequencies of neutral alleles should be uncorrelated in reproductively isolated populations, and to predictions regarding rates at which frequencies of neutral alleles are expected to diverge among such populations. The slender wild oat species, *Avena barbata*, is exceptionally well suited for experimental tests of these two predictions. This species was introduced into California during the Spanish period, and it has since become established as a prominent component of grassland and oak savannah communities throughout the state (11). Thus, the maximum time that any local population has existed cannot exceed 400 years, and the composition of the ancestral gene pool from which the California populations orginated can also be inferred from Mediterranean collections of the species. We have estimated allelic frequencies for five enzyme loci in 14 local populations of *A. barbata* from different ecogeographical regions of California to determine how much genetic divergence has occurred since this species was introduced to the state. We have also assayed nine collections of *A. barbata* from the Mediterranean Basin to determine the composition of the ancestral gene pool. The results establish, contrary to predictions of the neutral model, that the distribution of genetic variability at these loci in California is nonrandom and is strongly correlated with environment. Rates of differentiation among populations are also not in accord with predictions based on the drift of neutral alleles. Our analysis of the results leads us to the conclusion that natural selection, operating in different ways in different environments, is the main factor responsible for the observed patterns of genic variability in *A. barbata*.

MATERIALS AND METHODS

The sampling strategy adopted in California was to select populations for study at intervals of about 120 km on a north–south transect from San Diego on the south to Redding on the north (see Fig. 1). A rough east–west transect, about 80 km north of San Francisco, was also obtained that began at Bodega Bay on the Pacific coast and terminated near Fiddletown in the foothills of the Sierra Nevada Mountains. About 150 panicles containing mature seeds were selected at random from each population. When the population occupied a site 2500 m^2 or smaller in area, the entire population was sampled. If the population extended over a larger area, sampling was restricted to a fairly central region of about 2000 m^2. All populations sampled were large, the number of plants within the sampling area exceeding 100,000 in all cases. No populations occupying sites that showed evidence of disturbance by man or domestic animals were included in the study.

The inheritance and linkage relationships and the electrophoretic techniques used to detect the allozymes have been described (12). The five loci studied are: Esterases, three loci (E_4, E_9, E_{10}); phosphatase (EC 3.1.3.2), one locus (P_5); peroxidase (EC 1.11.1.7), one locus (APX_5). All allozyme assays were performed on tissue from seedlings grown from seeds collected in nature. The numbers of seedlings assayed per population, and the numbers of families (derived from single panicles) represented in each population are given in Table 1. In addition to the five enzyme polymorphisms, each

population was scored for two loci governing morphological polymorphisms, *hairy* against *nonhairy* lemma (*H*, *h*) and *black* against *grey* lemma (*B*, *b*).

RESULTS

Fig. 1 is a map of California on which the location of each population of this study is shown, and two major climatic zones are identified (13). Region I is the Mediterranean warm summer zone (mean temperature of the warmest month exceeds 22°C) and Region II the Mediterranean cool summer zone (mean temperature of the warmest month below 22°C). In addition to mean temperature, these two regions differ in rainfall and in many other climatic factors; they also differ topographically. Region I, which includes the extensive semiarid grasslands and oak savannah of the foothills bordering the central valley, is much more uniform environmentally than Region II, which includes the intermontane regions of the coastal strip and higher foothills of the Sierra Nevada mountains, and features marked diversity of microenvironments that vary from mesic to semiarid. The distinction between these two regions is important, because previous investigations have shown that populations polymorphic for morphological markers are almost entirely restricted to Region I (14).

Allelic frequencies for each of the enzyme loci, together with frequencies of the two morphological morphs, are given in Table 1. These frequency data show that eight of the nine populations in Region I (Populations 1, 2, 3, 5, 6, 13, 15, 16) are monomorphic at all loci, and that the remaining population in this region (Population 4) is monomorphic at four enzyme loci and virtually monomorphic at the fifth locus. Moreover, these populations are monomorphic for the same allele at each enzyme locus, and are also fixed for the *black* and *hairy* lemma morphs. In contrast, most of the Region II populations are highly polymorphic [the frequency data for the Calistoga and San Rafael populations are taken from an earlier study (15)]. The only exception is the Lick 2 population, which is monomorphic at all loci. Interestingly, with the exception of the *H*, *h* locus, the allele fixed in the Lick 2 population is the opposite of the one fixed in the warmer, more arid sites of Region I.

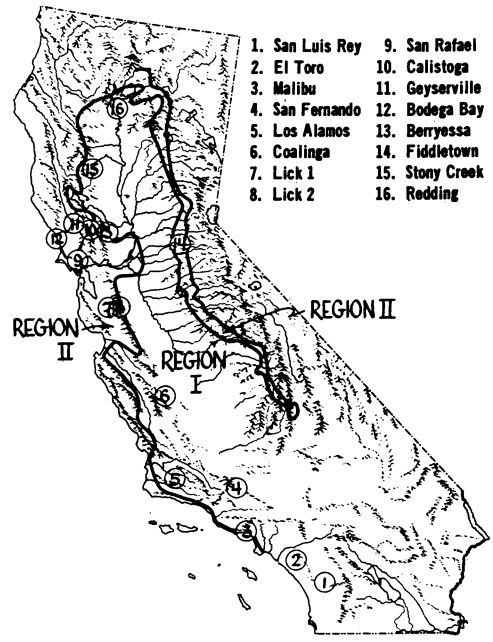

FIG. 1. Location of sampling sites in California.

To obtain quantitative measures of the number of gene differences among the California populations, we have computed the genetic identity and genetic distance statistics proposed by Nei (10). Genetic identity is given by

$$I = J_{xy}/\sqrt{J_x J_y},$$

where $J_{xy} = \Sigma x_i y_i$, $J_x = \Sigma x_i^2$, $J_y = \Sigma y_i^2$, and x_i and y_i denote the frequency of the i^{th} allele in population X and Y, respectively. The arithmetic mean of J_{xy}, J_x and J_y over loci

TABLE 1. *Allelic frequencies of the most common allele for five enzyme loci and two loci governing morphological characters in 16 California populations of A. barbata*

Population	*N*	*n*	Allozyme or morph frequency						
			$E_4^{(1)}$	$E_9^{(2)}$	$E_{10}^{(2)}$	$P_5^{(2)}$	$APX_5^{(1)}$	*B*	*H*
1. San Luis Rey	68	360	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2. El Toro	81	81	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3. Malibu	50	279	1.00	1.00	1.00	1.00	1.00	1.00	1.00
4. San Fernando	84	84	0.97	1.00	1.00	1.00	1.00	1.00	1.00
5. Los Alamos	59	310	1.00	1.00	1.00	1.00	1.00	1.00	1.00
6. Coalinga	47	262	1.00	1.00	1.00	1.00	1.00	1.00	1.00
7. Lick 1	38	38	0.00	0.00	0.19	0.18	0.19	0.07	1.00
8. Lick 2	72	399	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9. San Rafael	87	740	1.00	—	0.00	1.00	0.89	—	—
10. Calistoga	54	460	0.36	—	0.47	0.52	0.54	—	—
11. Geyserville	—	211	0.30	0.77	0.80	1.00	0.00	0.20	0.89
12. Bodega Bay	73	535	0.86	0.92	0.98	1.00	0.82	0.81	0.99
13. Berryessa	75	75	1.00	1.00	1.00	1.00	1.00	1.00	1.00
14. Fiddletown	62	339	1.00	1.00	0.83	1.00	1.00	1.00	1.00
15. Stony Creek	66	330	1.00	1.00	1.00	1.00	1.00	1.00	1.00
16. Redding	71	373	1.00	1.00	1.00	1.00	1.00	1.00	1.00

N = number of families (derived from a single panicle) assayed per population; *n* = number of seedlings assayed per population.

TABLE 2. *Genetic identity (above diagonal) and genetic distance (below diagonal) values for 16 California populations of A. barbata*

Population 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1.000	1.000	1.000	1.000	1.000	0.281	0.167	0.653	0.741	0.609	0.991	1.000	0.996	1.000	1.000
2	0.000	1.000	1.000	1.000	1.000	0.281	0.167	0.653	0.741	0.609	0.991	1.000	0.996	1.000	1.000
3	0.000	0.000	1.000	1.000	1.000	0.281	0.167	0.653	0.741	0.609	0.991	1.000	0.996	1.000	1.000
4	0.000	0.000	0.000	1.000	1.000	0.281	0.172	0.653	0.747	0.614	0.992	1.000	0.995	1.000	1.000
5	0.000	0.000	0.000	0.000	1.000	0.281	0.167	0.650	0.741	0.609	0.991	1.000	0.996	1.000	1.000
6	0.000	0.000	0.000	0.000	0.000	0.281	0.167	0.653	0.741	0.609	0.991	1.000	0.996	1.000	1.000
7	1.269	1.269	1.269	1.245	1.269	1.269	0.985	0.318	0.884	0.350	0.323	0.281	0.307	0.281	0.281
8	1.792	1.792	1.792	1.758	1.792	1.792	0.015	0.253	0.797	0.264	0.205	0.167	0.199	0.167	0.167
9	0.427	0.427	0.427	0.431	0.427	0.427	1.147	1.373	0.566	0.493	0.658	0.653	0.708	0.653	0.653
10	0.300	0.300	0.300	0.292	0.300	0.300	0.124	0.227	0.570	0.533	0.759	0.741	0.765	0.741	0.741
11	0.496	0.496	0.496	0.487	0.496	0.496	1.051	1.330	0.707	0.630	0.706	0.609	0.604	0.609	0.609
12	0.009	0.009	0.009	0.008	0.009	0.009	1.129	1.586	0.418	0.276	0.348	0.991	0.985	0.991	0.991
13	0.000	0.000	0.000	0.000	0.000	0.000	1.269	1.792	0.427	0.300	0.496	0.009	0.996	1.000	1.000
14	0.004	0.004	0.004	0.005	0.004	0.004	1.180	1.614	0.346	0.268	0.504	0.015	0.004	0.996	0.996
15	0.000	0.000	0.000	0.000	0.000	0.000	1.269	1.792	0.427	0.300	0.496	0.009	0.000	0.004	1.000
16	0.000	0.000	0.000	0.000	0.000	0.000	1.269	1.792	0.427	0.300	0.496	0.009	0.000	0.004	0.000

is used to calculate I when several loci are considered simultaneously. Genetic distance between pairs of populations is defined as $D = -\log_e I$. Table 2 gives genetic identity and genetic distance values for the enzyme loci among all pairs of California populations. These values show that the nine populations in Region I are identical ($I = 1.000$, $D = 0.000$) to each other (the limited polymorphism at one locus in the San Fernando population does not affect I and D to an accuracy of three decimal places). In contrast, the populations of Region II, with the exception of Lick 1 and 2, are all highly differentiated from each other ($\bar{I} = 0.540$ for the region). Thus, the genetic identity and genetic distance measures also bring out the remarkable and nonrandom pattern in which genetic differentiation has occurred in California.

To ascertain how representative the California populations are of the ancestral gene pool, we have estimated allelic frequencies in collections of *A. barbata* from nine areas of the Mediterranean Basin*. The results, given in Table 3, show that the same alleles occur in both gene pools with the exception of $APX_5^{(1)}$, $P_5^{(1)}$ and h, which occur in California, but were not found in our samples from the Mediterranean, and $APX_5^{(4)}$, $APX_5^{(5)}$, and $P_5^{(3)}$, which occur in the Mediterranean, but have not been observed in California. To obtain a quantitative measure of genetic identity within and between the two gene pools, we have calculated Nei's genetic identity statistics within and between and California and Mediterranean samples. The mean genetic identity over all pairs of California populations is 0.715 ± 0.028, and that for the Mediterranean populations is 0.666 ± 0.031, while mean genetic identity between the California and Mediterranean populations is 0.726 ± 0.029. This result indicates that the California gene pool does not differ greatly in overall composition from its ancestral gene pool, a not surprising finding since botanical history shows that multiple introductions of *A. barbata* occurred to California from the Mediterranean area. Yet, at the same time patterns of differentiation unique to California have developed. In particular, the extensive monomorphism in Region I, featuring identical alleles at five enzyme loci and two loci governing morphological characters, finds no parallel in the Mediterranean.

Given that the California gene pool is representative of the ancestral gene pool, we can ask whether differentiation in California has proceeded at rates that are consistent with the predictions of the neutral model. Nei has shown that genetic distance $D = 2\alpha t - \log_e I_0$, approximately. In this equation, α is the rate of allelic substitution per locus per year (which equals mutation rate per locus per year for neutral alleles), t is time in years since separation, and I_0 is the initial genetic identity between populations. The value of I_0 is expected to be near unity in most cases, so that the second term in the above equation can generally be neglected. A reasonable estimate of time (t) since the establishment of the majority of California populations is 150 years, because most of the colonization of California by *A. barbata* occurred during the first half of the nineteenth century. In a study of mutation rates for enzyme loci in *Hordeum vulgare*, another grass species, Kahler and Allard (unpublished data) have found no electrophoretically detectable mutants in more than 8×10^5 gametes examined; if we extrapolate this result to *A. barbata*, it seems likely that $\alpha < 10^{-5}$. Thus, the best estimates of the relevant parameters are $I_0 = 1$, $t = 150$ years, and $\alpha < 10^{-5}$.

Under the above assumptions regarding I_0, t, and α, the

* We are indebted to Dr. J. C. Craddock, Crops Research Division, Agricultural Research Service, U.S. Department of Agriculture for making these collections available to us.

TABLE 3. *Distribution of enzymatic and morphological variability in Mediterranean collections of A. barbata, reported as the number of collections exhibiting each allele at each locus*

	Enzyme locus											Morph								
	E_4		E_9		E_{10}		APX_5					P_5								
Location	1*	2	1	2	1	2	1	2	3	4	5	1	2	3	B	b	H	h	Number of collections	Total number of plants observed
Sicily	1	1	1	1		2		2					2		2		2		2	12
Israel	16	4	7	12	8	12		7	6		7		19		19		19		19	108
Turkey	2	5	1	5		6		2	5				6		6		6		6	35
Greece	1			1		1				1			1		1		1		1	5
Sardinia	3	5	3	5	6	3		1	4	2	1		8		7	2	8		8	43
Algeria	5	2	1	6	3	4		6		1			7		7		7		7	42
Crete	1		1			1			1					1	1	1	2		2	4
Italy		5	4	2	3	3			2	4			5		5		5		5	32
Corsica	1			1	1			1					1		1		1		1	6

* The number denotes the allozyme whose frequency is tabulated.

exceptional uniformity of Region I ($\bar{D}$ slightly larger than zero as a result of minor polymorphism at one locus in the San Fernando population) is expected under the neutral model, provided it is assumed that all introductions to this region included only one genotype, specifically the single seven-locus genotype characteristic of the region. This assumption appears to be untenable, however, on two counts. First, if we consider that the California gene pool as a whole received a near random sample of genetic variability from the Mediterranean, it seems highly unlikely that only a single genotype, particularly one that apparently does not occur in the ancestral gene pool, was introduced into Region I. Second, it is known that *A. barbata* has frequently been transported between Regions I and II as a result of agricultural activities, so providing recurring opportunities for migration from the diversity of Region II into Region I. Such migration is also at variance with the assumption that no more than one genotype has been introduced into Region I.

When values of α are computed for Region II, with the same assumptions as above that $I_0 = 1$ and $t = 150$ years, nearly all estimates for pairs of populations exceed 1×10^{-3}, and $\bar{\alpha}$ for the region is 2.4×10^{-3}. However, the value of I_0 may be smaller than unity in some cases due to founder effects; it is also possible that some of the populations studied have been isolated reproductively for more than 150 years. To obtain minimum estimates of α we have assumed: (*i*) that $N_x = N_y = 1$ at time of divergence; (*ii*) that $J_{xy} = 0.6$, so that $I_0 = 0.75$; and (*iii*) that $t = 400$ years. Even under these extreme assumptions, values of α are greater than 1×10^{-3} for the majority of population pairs in the region, i.e., at least 100 times larger than the measured mutation rate for electrophoretically detectable variants in grasses. This result, thus, points strongly to the conclusion that divergence in Region II has been more rapid than is predicted by the neutral model, even when allowance is made for possible effects of the small number of loci studied on the estimates of α.

DISCUSSION

The distribution of allozyme frequencies of *A. barbata* in California is difficult to reconcile with the hypothesis that electrophoretically detectable variants are adaptively neutral. Region I is too uniform and Region II is too highly differentiated for their genetic patterns to have been established by sampling accidents. Explanations that attempt to account for the distribution of genetic variability in California on the basis of the drift on neutral alleles become highly convoluted, because they must take into account both the contrasting patterns of differentiation in the two regions and also the fact that the California gene pool represents a nearly random sample from the ancestral Mediterranean gene pool of this species. In particular, such explanations require the assumption that migration was on a random basis into one region and on a nonrandom basis into the other region, an assumption that is at variance with botanical history. Such explanations also require the biologically unreasonable assumption that mutation rates differ in the two regions.

Neo-Darwinian evolutionary theory appears to offer a more economical explanation of the observations. *A. barbata* was introduced through numerous colonizing episodes that brought to California a representative sample of the rich Mediterranean gene pool (11). The frequent migrations from place to place that characterize the population biology of this species in California allowed repeated tests of the adaptiveness of numerous gene combinations in the full range of new environments available to the species. During this period of genetic experimentation, a genotype appeared that was highly adapted to the warmer more arid habitats; it soon became almost the exclusive genotype in Region I. Another genotype highly adapted to mesic habitats also evolved that became predominant in the cooler moister areas of Region II, such as the Lick 2 site. In a study of the spatial distribution of gene frequencies within environmentally heterogeneous sites in Region II, Hamrick and Allard (unpublished results) found that extreme differentiation occurs over very short distances. They also found parallelisms between gene frequencies and environment that featured monomorphism for the genotypic combination of Region I in the most xeric areas of their sites, monomorphism for the mesic (Lick 2) genotypic combination in the most mesic parts of their sites, and polymorphism in which gene frequencies were correlated with degree of aridity in environmentally intermediate subsites. This remarkable correspondence between gene frequency and environment on both macro- and microgeographical scales leads us to the conclusion that selection played a major role in molding the random sample of genes received from the

ancestral gene pool into the unique patterns that developed so rapidly in California.

What maintains the polymorphisms observed in the great majority of populations in Region II? Studies of several populations (15) have revealed a consistent excess of heterozygotes over expectations based on the assumption of neutrality, leading to the conclusion that the polymorphisms are maintained by some sort of balancing selection. Such selection was also revealed by further analyses of the data of the present study; upon dissection of this selection into viability and fertility components, we found that the selective differential results almost entirely from viability rather than from fertility selection (unpublished results). In a detailed study of multilocus allozyme polymorphisms in *A. barbata*, Babbel and Allard (unpublished results) have found nonrandom associations among alleles at different loci of types that lead to balancing selection and have concluded that epistatic interactions among alleles at different loci play an important role in the maintenance of genetic variability in *A. barbata*.

This work was supported in part by NIH Grant GM 10476 and NSF Grant GB 13213.

1. Lewontin, R. C. & Hubby, J. L. (1966) *Genetics* **54,** 595–604.
2. Selander, R. K., Hunt, W. G. & Yang, S. Y. (1969) *Evolution* **23,** 379–390.
3. Harris, H. (1966) *Proc. Roy. Soc. Ser.* B, **164,** 298–310.
4. Allard, R. W. & Kahler, A. L. (1971) in *Stadler Symposia* (Univ. of Missouri), Vol. 3, pp. 9–24.
5. Ayala, F. J., Powell, J. R. & Dobzhansky, Th. (1971) *Proc. Nat. Acad. Sci. USA* **68,** 2480–2483.
6. Kimura, M. (1968) *Nature* **217,** 624–626.
7. King, J. L. & Jukes, T. E. (1969) *Science* **164,** 788–798.
8. Jukes, T. E. (1965) *Amer. Sci.* **53,** 477–487.
9. Maynard Smith, J. (1970) *Amer. Natur.* **104,** 231–237.
10. Nei, M., *Amer. Natur.*, in press.
11. Robbins, W. W. (1940) *Univ. Calif. Agric. Exp. Stat. Bull.* No. 637.
12. Marshall, D. R. & Allard, R. W. (1969) *J. Hered.* **60,** 17–19.
13. Durrenberger, R. W. (1960) *Patterns on the Land* (Roberts Publishing Co., Northridge, Calif.).
14. Marshall, D. R. & Jain, S. K. (1969) *Nature* **221,** 276–278.
15. Marshall, D. R. & Allard, R. W. (1970) *Genetics* **66,** 393–399